Nonparametric regression – some approaches*

Tomislav Marošević†

Abstract. *In this paper we describe two approaches to nonparametric regression. First, we consider the nearest neighbour approach, as a procedure which serves mainly for obtaining an* ad hoc *smoothing and interpolating. Next, we describe the roughness penalty approach. This gives a certain compromise between the demand for goodness-of-fit of regression curve to the given data and the condition that the regression curve has not too many oscillations.*

Key words: *nonparametric regression, nearest neighbour approach, roughness penalty approach, penalized least squares, smoothing splines*

Sažetak. **Neparametarska regresija — neki pristupi.** *U ovom radu opisana su dva pristupa problemu neparametarske regresije. Razmotren je pristup najbližeg susjeda, kao postupak koji služi uglavnom za dobivanje* ad hoc *glađenja i interpolatora. Također, opisan je i pristup nametanjem krivudavosti. Njime se postiže određena mjera između zahtjeva za dobrom prilagođenosti funkcije regresije danim podacima i uvjeta da pripadna krivulja regresije nema prevelike oscilacije.*

Ključne riječi: *neparametarska regresija, pristup najbližeg susjeda, pristup nametanjem krivudavosti, najmanji kvadrati s nametom, glađeni spline-ovi*

1. Introduction

For the given data (t_i, y_i), $i = 1, \ldots, m$, obtained by a certain experimental or empirical way, it is often necessary to explore relations between independent and dependent variables which the data represent.

*The lecture presented at the Mathematical Colloquium in Osijek organized by Croatian Mathematical Society - Division Osijek, December 1, 1995.

†Faculty of Electrical Engineering, Department of Mathematics, Istarska 3, HR-31 000 Osijek, Croatia, e-mail: tommar@osijek.etfos.hr

As a procedure of analyzing these interrelations which are described as "cause and effect", the regression has two main purposes:
1) to explore and to present the relationship between the design variable t (respectively p independent variables $\mathbf{x} = [x_1, \ldots, x_p]^T$ in the p-dimensional case) and the response variable y;
2) to predict, for any given point t, the values of observation y at the point t.

Assume that t_i and y_i are related by the regression model:

$$y_i = g(t_i) + \varepsilon_i, \quad i = 1, \ldots, m,$$

where ε_i represent errors with mean zero ($E\varepsilon_i = 0$) and common variance σ^2 ($\operatorname{Var}\varepsilon_i = \sigma^2$), and $g(t_i)$ are values of some unknown function g at the points $t_1, \ldots, t_m$. The function g is usually referred to as the regression function.

Parametric regression models assume that the form of g is known, except for finitely many unknown parameters (see [**?**], [**?**]). For instance, the following models are well known:
a) linear regression of the form $g(t) = at + b$, where a, b are parameters;
b) nonlinear exponential model $g(t) = a + b \cdot e^{ct}$, where $a,\ b,\ c$ are parameters.
Parameters in a model are estimated on the basis of the given data by some appropriate method. If one can assume that the data are contaminated by errors with normal distributions $N(0, \sigma^2)$, then the least squares method will be used, that is, a discrete L_2 approximation (see [**?**]). In applications L_1 or L_∞ approximations are often used, too (see [**?**]).

A nonparametric regression model allows a much greater level of indefiniteness and generally only assumes that g belongs to some infinite dimensional space of functions (see [**?**]).

In *Section 2.* we describe the nonparametric regression by the *nearest neighbour approach*, and in *Section 3.* we consider the *roughness penalty approach*.

2. Nearest neighbour approach

General property of this approach is that regression is used in a more or less *ad hoc* fashion, without much thought as to the mechanism underlying the system under consideration (see [**?**]).

A nearest neighbour approach is a simple means for smoothing the data, which represents a very local procedure of digital filtering.

Among the simplest of the digital filters is the **hanning window**, at which the equal distance between knots in t-domain is supposed. By using such window, each observation $y(t)$ is replaced by the average according to the rule

$$y(t) \leftarrow \frac{y(t-1) + 2y(t) + y(t+1)}{4}.$$

A hanning filter is excellent for "smoothing out rough edges". But, it is much less satisfactory for neglecting those observations which deviate much from the

others. The attempt to correct strongly deviating observations requires the use of the H filter several times, but it can disturb the internal structure of the data.

Example 1. *Repeated application of the hanning window in the following case essentially disturbs an internal structure of the data.*

t	1	2	3	4	5	6	7	8	9
$y(t)$	1	1	1	1	1000	1	1	1	1
H	1	1	1	250.75	500.5	250.75	1	1	1
HH	1	1	63.44	250.75	375.62	250.75	63.44	1	1
HHH	1	16.61	94.66	235.14	313.18	235.14	94.66	16.61	1

Continuation of this procedure would lead to stationary data.

An alternative filter, which eliminates the major effects of very deviating observations ("outliers") and prevents their expanding through the entire data set, is obtained by the use of the so called **median smooth algorithm**. It is applied by the rule

$$y(t) \leftarrow \operatorname{med}\{y(t-1), y(t), y(t+1)\},$$

where med denotes the middle element according to the magnitude among the given three neighbour elements.

Repeated applications of this algorithm are not going to obscure the data. Already after two or three iterations we obtain the resulting transformed set, which is not changed by further smoothing (the label of the last repetition after which there are no changes is 3R). The 3R filter can be used to eliminate the effect of the outliers (i.e. very much deviating observations), but it usually induces the rough edges. Hence, the H filter can be applied to smooth away the rough edges.

Example 2. *Most of the good properties of the combined 3RH procedure can be illustrated in the case of the data set (consisting of 21 days of returned goods at a large department store, see [**?**]) given in* Table *1. For the initial and final point (day) we follow the convention of not changing the values of the endpoints for the 3R filter, and for the H filter we take the average of the end value and its first neighbour value.*

Figure 1: Nearest neighbour approach

Day	1	2	3	4	5	6	7	8	9	10	11
Returns	30	42	53	62	33	68	72	81	50	75	62
3	30	42	53	53	62	68	72	72	75	62	62
3R									72		
3RH	36.0	41.75	50.25	55.25	61.25	67.5	71.0	72.0	69.5	64.5	62
Day	12	13	14	15	16	17	18	19	20	21	.
Returns	51	80	60	51	25	44	41	50	57	63	.
3	62	60	60	51	44	41	44	50	57	63	.
3R						44					.
3RH	61.5	60.5	57.75	51.5	45.75	44.0	45.5	50.25	56.75	60.0	.

Table 1

The use of such procedures as 3RH is mainly to provide an *ad hoc* smoothing and interpolation.

3. Roughness penalty approach

If our aim is to minimize the deviation (i.e. residual error) in the choice of a regression curve, the solution could be any regression curve which interpolates the given data points. For instance, joining these points by straight lines would give a polygonal regression. If we also add condition of smoothness for the interpolating curve, the obtained curve could have a lot of variations, in accordance with the variations of the given data.

Quantifying the roughness of a curve can be done in various ways. For instance, one could consider $\max_{t\in[a,b]} |g''(t)|$, or the number of inflection points. As a global measure of roughness on an interval $[a, b]$ (where $a < t_1 < \ldots < t_m < b$), one uses the integrated squared second derivative $\int_a^b [g''(t)]^2 dt$.

Suppose that $g \in C^2([a,b])$ and let us define the penalized sum of squares

$$F(g) = \sum_{i=1}^{m} [y_i - g(t_i)]^2 + \alpha \cdot \int_a^b [g''(t)]^2 dt\,, \tag{1}$$

where $\alpha \in [0, \infty\rangle$ is the given smoothing parameter.

The regression function $\hat{g}$ is obtained by minimizing the functional F over the class of all continuously twice-differentiable functions g, $g \in C^2([a,b])$. The smoothing parameter α represents the relation between the residual error $\sum_{i=1}^{m} [y_i - g(t_i)]^2$ and the roughness $\int_a^b [g''(t)]^2 dt$. If α is large, the roughness of the minimizer $\hat{g}$ of the functional F is small, and conversely, the roughness of $\hat{g}$ is large, provided α is small.

One can show that the solution $\hat{g}$ to the minimization problem for the functional F is a natural cubic spline for the data $(t_i, \hat{g}(t_i))$.

Definition 1. *Suppose we are given real numbers $t_1, \ldots, t_m$ on some interval $[a,b]$ satisfying $a < t_1 < \ldots < t_m < b$. A function g defined on $[a,b]$ is a cubic spline if the following two conditions are satisfied:*

i) on each of the intervals $\langle a, t_1\rangle, \langle t_1, t_2\rangle, \ldots, \langle t_m, b\rangle$, g is a cubic polynomial;

ii) the polynomial pieces fit together at the points t_i in such a way that g itself and its first and second derivatives are continuous at each knot t_i, and hence on the whole of $[a,b]$.

Denote $g_i = g(t_i)$, $\gamma_i = g''(t_i)$, $i = 1, \ldots, m$. Then *a natural cubic spline g* (for which $\gamma_1 = \gamma_m = 0$ by definition) is completely specified by the vectors (see [**?**])

$$\mathbf{g} = [g_1, \ldots, g_m]^T, \quad \boldsymbol{\gamma} = [\gamma_2, \ldots, \gamma_{m-1}]^T.$$

Let us state some theorems, the proofs of which can be found in [**?**].

Theorem 1. *The vectors $\mathbf{g}$ i $\boldsymbol{\gamma}$ specify a natural cubic spline if and only if the condition*

$$Q^T \mathbf{g} = R\boldsymbol{\gamma} \tag{2}$$

*is satisfied. If (**??**) is satisfied, then the roughness penalty will satisfy*

$$\int_a^b [g''(t)]^2 dt = \boldsymbol{\gamma}^T R \boldsymbol{\gamma} = \mathbf{g}^T K \mathbf{g}, \quad K = Q R^{-1} Q^T, \tag{3}$$

where $h_i = t_{i+1} - t_i$, $i = 1, \ldots, m-1$, and

$$Q = \begin{bmatrix} \frac{1}{h_1} & 0 & \cdots & 0 \\ -(\frac{1}{h_1}+\frac{1}{h_2}) & \frac{1}{h_2} & \ddots & \vdots \\ \frac{1}{h_2} & -(\frac{1}{h_2}+\frac{1}{h_3}) & \ddots & 0 \\ 0 & \frac{1}{h_3} & \ddots & \frac{1}{h_{m-2}} \\ \vdots & \ddots & \ddots & -(\frac{1}{h_{m-2}}+\frac{1}{h_{m-1}}) \\ 0 & \cdots & 0 & \frac{1}{h_{m-1}} \end{bmatrix},$$

$$R = \begin{bmatrix} 2(h_1+h_2) & h_2 & 0 & \cdots & 0 \\ h_2 & 2(h_2+h_3) & h_3 & \cdots & \vdots \\ 0 & h_3 & 2(h_3+h_4) & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & h_{m-2} \\ 0 & \cdots & 0 & h_{m-2} & 2(h_{m-2}+h_{m-1}) \end{bmatrix}.$$

Theorem 2. *Suppose $m \geq 2$ and $t_1 < t_2 < \ldots < t_m$. Given any values $z_1, \ldots, z_m$, there is a unique natural cubic spline g with knots at the points t_i satisfying $g(t_i) = z_i$, for $i = 1, \ldots, m$.*

Theorem 3. *Suppose $m \geq 2$, and let g be the natural cubic spline interpolant to the values $z_1, \ldots, z_m$ at points $t_1 < t_2 < \ldots < t_m$, satisfying $a < t_1 < t_2 < \ldots < t_m < b$. Let $\overline{g}$ be any function in $C^2([a,b])$ for which $\overline{g}(t_i) = z_i$, $i = 1, \ldots, m$. Then*

$$\int_a^b [g''(t)]^2 dt \leq \int_a^b [\overline{g}''(t)]^2 dt,$$

where the equality holds only if $\overline{g}$ and g are identical.

Theorem 4. *Suppose $m \geq 3$ and let $t_1 < t_2 < \ldots < t_m$ be points satisfying $a < t_1 < t_2 < \ldots < t_m < b$. Given the data points $y_1, \ldots, y_m$ and a strictly positive smoothing parameter α, let $\hat{g}$ be the natural cubic spline with knots at the points $t_1, \ldots, t_m$, for which $\hat{\mathbf{g}} = (I + \alpha K)^{-1}\mathbf{y}$.*

Then, for any $g \in C^2([a,b])$

$$F(\hat{g}) \leq F(g),$$

where the equality holds only if g and $\hat{g}$ are identical.

For $(t_i, y_i),\ i = 1, \ldots, m-1$ the smoothing natural cubic spline, as the solution to the problem (**??**), has the following form on the intervals $[t_j, t_{j+1}]$:

$$\hat{g}(t) = \begin{cases} \hat{g}_1 - (t_1 - t)\hat{g}'(t_1), & t \le t_1, \\ \frac{(t-t_j)\hat{g}_{j+1} + (t_{j+1}-t)\hat{g}_j}{h_j} - \frac{1}{6}(t-t_j)(t_{j+1}-t)\cdot\left[(1+\frac{t-t_j}{h_j})\gamma_{j+1} + (1+\frac{t_{j+1}-t}{h_j})\gamma_j\right], & t_j \le t \le t_{j+1}, \\ \hat{g}_m + (t-t_m)\hat{g}'(t_m), & t \ge t_m, \end{cases} \tag{4}$$

where $\hat{\mathbf{g}} = [\hat{g}_1, \ldots, \hat{g}_m]^T$ is the solution to the equation

$$(I + \alpha K)\mathbf{g} = \mathbf{y}, \quad K = QR^{-1}Q^T, \tag{5}$$

and the numbers $\gamma_2, \ldots, \gamma_{m-1}$ can be obtained by solving the equation

$$(R + \alpha Q^T Q)\boldsymbol{\gamma} = Q^T \mathbf{y}\,. \tag{6}$$

The equation (**??**) follows from (**??**) and (**??**).

Remark 1. *The smoothing parameter α can be chosen on the basis of a free subjective choice, or conversely, by using some choice method, as for instance the "cross-validation" (see [**?**]).*

Remark 2. *A more general problem of the weighted smoothing can be set up by minimizing the functional:*

$$F_W(g) = \sum_{i=1}^{m} w_i \left[y_i - g(t_i)\right]^2 + \alpha \cdot \int_a^b [g''(t)]^2 dt\,, \tag{7}$$

where $w_i > 0,\ i = 1, \ldots, m$, are the data weights. Analogously to the Theorem 4.*,*

$$\hat{\mathbf{g}} = (W + \alpha K)^{-1} W \mathbf{y}, \tag{8}$$

*where W is a diagonal matrix with diagonal elements w_i ([**?**]).*

Example 3. *For the artificial data given in* Table 2. *we have made the smoothing spline which results from the roughness penalty approach, with a free choice of the smoothing parameter α.*

t	0	2	3	5	8	11	12	16	20
$y(t)$	10	8	5	4	3.5	6	7	8	8.5

Table 2.

Calculations were done using Mathematica. *In* Figure 2 *we illustrate the graphs of the splines obtained for $\alpha = 0$ (interpolation), $\alpha = 10$, and $\alpha = 100$, respectively.*

Figure 2: Regression by penalized least squares

Remark 3. *There is a natural generalization of the smoothing splines in dimensions two or higher. In this case, some features of the fitting by one dimensional splines are carried over, and some are not ([**?**]).*

References

[1] R. L. Eubank, *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, Inc., 1988.

[2] P. J. Green, B. W. Silverman, *Nonparametric Regression and Generalized Linear Models*, Chapman&Hall, 1994.

[3] J. R. Rice, *Approximation of Functions, Vol. I.*, Addison–Wesley, Reading, 1964.

[4] G. J. S. Ross, *Nonlinear Estimation*, Springer-Verlag, 1990.

[5] J. R. Thompson, R. A. Tapia, *Nonparametric Function Estimation, Modeling and Simulation*, SIAM, Philadelphia, 1990.

[6] S. Wolfram, *Mathematica - A System for Doing Mathematics by Computer*, Addison-Wesley, 1991.